Exercise 21B

1. For each set of data below calculate the standard deviation:

(a) (i) 19.0, 23.4, 36.2, 18.7, 15.7

(ii) 0.4, −1.3, 7.9, 8.4, −9.4

(b) (i) 28, 31, 54, 28, 17, 30

(ii) 60, 18, 42, 113, 95, 23

2. For each set of data below calculate the standard deviation:

(a) 1, 1, 2, 3, 5 (b) 3, −2, 4, −2, 5, 2

EXAM HINT

If you study the Statistics option, you will see that we use μ for the mean and Σ for the standard deviation when we work with whole populations. In the core, all samples consist of the whole population, so the symbols can be used interchangeably. μ and Σ are used in the formula booklet.

3. The ordered set of data 5, 5, 7, 8, 9, x, 13 has interquartile range equal to 7.

(a) Find the value of x.

(b) Find the standard deviation of the data set. *[5 marks]*

4. Consider the five numbers, 2, 5, 9, x and y. The mean of the numbers is 5 and the variance is 6. Find the value of xy. *[7 marks]*

5. In five tests Suewan has an average of 23 marks and a standard deviation of 4 marks. In her sixth test she scores 32. What is the overall standard deviation of her marks? *[7 marks]*

6. The mean of a set of 15 data items is 600 and the standard deviation is 12. Another piece of data is discovered and the new mean is 600.25. What is the new standard deviation? *[7 marks]*

7. If the sum of 20 pieces of data is 1542 find the smallest possible value of $\sum x_i^2$. *[4 marks]*

8. (a) Explain why for any set of data $x_i - \bar{x}$ is no greater than the range.

(b) By considering the formula $s_n = \sqrt{\sum_i \frac{(x_i - \bar{x})^2}{n}}$, prove that the standard deviation is always less than or equal to the range. *[4 marks]*

21C Frequency tables and grouped data

It is common to summarise a large quantity of data in a **frequency distribution** table. This is a list of all the values the data takes, along with how often they occur. We could convert this into a list of all the data values and calculate the statistics as we had before, but it is enough to just imagine writing out a list, 16 ones, two twos, etc.

Worked example 21.3

Find the mean number of passengers observed in cars as they passed a school.

Passengers	Frequency
0	32
1	16
2	2
3 or more	0

(none in the first group, 16 in the second group 2, and 4 in the third group)

total number of passengers =

$(32 \times 0) + (16 \times 1) + (2 \times 2) + 0 = 20$

$\text{mean} = \frac{20}{50} = 0.4$

EXAM HINT

The mean does not have to be an achievable value; do not round to the nearest whole number.

This method suggests an important formula.

KEY POINT 21.3

Finding the mean from a frequency table:

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

where f_i is the frequency of the ith data value and $n = \sum_i f_i$ is the total number of data items.

We can work out $\overline{x^2}$ in a similar way, which gives the following formula for standard deviation.

KEY POINT 21.4

Standard deviation from a frequency table:

$$s_n^2 = \frac{\sum_i x_i^2 f_i}{n} - \bar{x}^2$$

EXAM HINT

The notation [a,c[means the same as $a < x < c$. You might also see this written as [a, c).

In Worked example 21.3, we knew the exact data values, but when we are dealing with grouped data, we do not have this level of precision. In order to work out the mean and standard deviation, our best and simplest assumption is that all the original values in a particular group are located at the centre of the group, called the mid-interval value. To find the centre of the group we take the mean of the largest and the smallest possible values in the group, called the upper and lower interval boundaries.

Worked example 21.4

Find the mean and standard deviation of the weight of eggs produced by a chicken farm. Explain why these answers are only estimates.

Weight of eggs, in g	Frequency
[100, 120[	26
[120, 140[	52
[140, 160[	84
[160, 180[	60
[180, 200[	12

Make a table using the mid-interval value for each group

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
110	26	2860	314 600
130	52	6760	878 800
150	84	12 600	1 890 000
170	60	10 200	1 734 000
190	12	2280	433 200
Sum:	234	34 700	5 250 600

Apply the formulae

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{34700}{234} = 148.3\,g\ (3SF)$$

$$s_n^2 = \frac{\sum_i x_i^2 f_i}{n} - \bar{x}^2 = \frac{5250600}{234} - 148.3^2 = 448.4$$

Therefore $s_n = 21.2\,g\ (3SF)$

These answers are only estimates because we have assumed that all the values in each group are at the centre, rather than using the actual data.

EXAM HINT

Whenever you find a mean or a standard deviation it is always worth checking that the numbers make sense in context. Given the data, an average of about 150 g here seems reasonable.

Sometimes the endpoints of the intervals shown in the table are not the actual smallest and largest possible values in that group. For example, when measuring length in centimetres it is common to round the values to the nearest integer, so 10–15 actually means [9.5, 15.5]. To find the mid-interval values we must first identify the actual interval boundaries.

Worked example 21.5

Estimate the mean of this data:

Age	Frequency
10 to 12	27
13 to 15	44
16 to 19	29

Carefully decide on the upper and lower interval boundaries. There should be no 'gaps' between the groups, because age is continuous data. You are 12 years old until your 13th birthday

Group	x_i	f_i	$x_i f_i$
[10, 13[	11.5	27	310.5
[13, 16[	14.5	44	638
[16, 20[	18	29	522
Sum:		100	1470.5

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{1470.5}{100} = 14.7 \text{ (3SF)}$$

Exercise 21C

1. Calculate the mean and standard deviation of each data set:

(a)

x	Frequency
0	16
1	22
2	8
3	4
4	0

(b)

x	Frequency
−1	10
0	8
1	5
2	1
3	1

2. Calculate the mean and standard deviation for each data set:

(a)

x	Frequency
10	7
12	19
14	2
16	0
18	2

(b)

x	Frequency
0.1	16
0.2	15
0.3	12
0.4	9
0.5	8

3. A group is described as '17 – 20'. State the upper and lower boundaries of this group if it is measuring:

(a) age in completed years

(b) number of pencils

(c) length of a worm to the nearest centimetre

(d) hourly earnings, rounded up to whole dollars.

4. Find the mean and standard deviation of each of the following sets of data:

(a) (i) x is the time taken to complete a puzzle in seconds

x	Frequency
[0,15[	19
[15,30[	15
[30,45[	7
[45,60[	5
[60,90[	4

(ii) x is the weight of plants in grams

x	Frequency
[50,100[	17
[100,200[	23
[200,300[	42
[300,500[	21
[500,1000[	5

(b) (i) x is the length of fossils found in a geological dig, to the nearest centimetre

x	Frequency
0 to 4	71
5 to 10	43
11 to 15	22
16 to 30	6

(ii) x is the power consumption of light bulbs, to the nearest watt

x	Frequency
90 to 95	17
96 to 100	23
101 to 105	42
106 to 110	21
111 to 120	5

(c) (i) x is the age of children in a hospital ward

x	Frequency
0 to 2	12
3 to 5	15
6 to 10	7
11 to 16	6
17 to 18	3

(ii) x is the amount of tips paid in a restaurant, rounded down to the nearest dollar

x	Frequency
0 to 5	17
6 to 10	29
11 to 20	44
21 to 30	16
31 to 50	8

5. In a sample of 50 boxes of eggs, the number of broken eggs per box is shown below:

Number of broken eggs	0	1	2	3	4	5	6
Number of boxes	17	8	7	7	6	5	0

(a) Calculate the median number of broken eggs per box.

(b) Calculate the mean number of broken eggs per box. *[4 marks]*

6. The mean of the data in the table is 32 and the variance is 136. Find the possible values of p and q.

x	Frequency
20	12
40	q
p	8

[8 marks]

Summary

- Most of statistics is based on trying to infer properties of a **population** based upon a **sample** from that population.
- To get a representative sample of the population, it is good practice to collect a **random sample**, where each member of the population is equally likely to be selected for the sample and the probability of selecting a member of the population is independent.
- An **outlier** is a correct but unusual data value.
- An **anomaly** is an unusual data value caused by a measurement error.
- **Discrete data** takes only a predefined value (it does not have to be an integer!).
- **Continuous data** can take any value in a given range. This type of data is generally grouped before we can work with it.
- **Standard deviation** (s_n) is a measure of how spread out the data is relative to the data's mean, and it takes into account all of the data.
- The square of the standard deviation is called the **variance** and it has the formula:

  $s_n^2 = \sum_i \frac{(x_i - \bar{x})^2}{n}$ or more commonly: $s_n^2 = \overline{x^2} - \bar{x}^2$
- Large datasets are summarised in **frequency distribution** tables and the mean and standard deviations can be calculated from these tables using the formulae:

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

where f_i is the frequency of the ith data value and $n = \sum_i f_i$ is the total number of data items,

and:

$$s_n^2 = \frac{\sum_i x_i^2 f_i}{n} - \bar{x}^2$$

- When investigating grouped data we must assume that every element has the mid-interval value of the group (the mean of the upper and lower boundaries).
- The methods for calculating statistics for grouped data vary slightly depending upon whether the data is discrete or continuous.

Introductory problem revisited

The magnetic dipole of an electron is measured in a very sensitive experiment 3 times. The values are 2.000 000 15, 2.000 000 12 and 2.000 000 9. Does this support the theory that the magnetic dipole is 2?

The average magnetic dipole is 2.000 000 12 which is pretty close to 2, but the standard deviation in the measurements is 0.000 000 245, and so the mean is approximately 5 sample standard deviations away from 2. Within the natural variation observed, the magnetic dipole cannot be said to be 2.

The difference between 2.000 001 2 and 2 might seem trivial, but it was this difference which inspired Richard Feynman to create a new theory of physics called Quantum electrodynamics, which did indeed predict this tiny difference from 2! This is an example of theory driving experiment which in turn creates new theory – the interplay between theoretical mathematics and reality.

Mixed examination practice 21

Short questions

1. A student takes the bus to school every morning. She records the length of the time, in minutes, she waits for the bus on 12 randomly chosen days. The data is summarised by:

$$\sum_{i=1}^{12} x_i = 49 \text{ and } \sum_{i=1}^{12} x_i^2 = 305.7$$

Calculate:

(a) the mean time she spends waiting for the bus

(b) the standard deviation of the times. *[5 marks]*

2. The average wavelength of light in nanometres emitted by a glowing wire is measured on 50 different occasions and the results are given below:

Wavelength in nm (λ)	Frequency
600–640	22
640–680	18
680–720	x
720–760	y

The mean of λ is calculated from this table as 653.6.

(a) Find the values of x and y.

(b) Calculate an estimate of the variance.

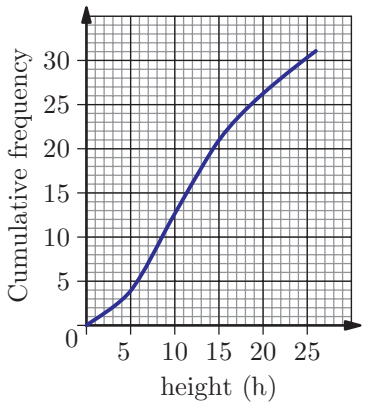
(c) Explain why this is only an estimate. *[9 marks]*

3. An experiment was conducted on the reaction times of 15 students (t) in seconds. The results were that the average reaction time was 0.2 s and the variance was 0.0025 s^2. A 16th student is observed later. She has a reaction time of 0.16 s. Find the new mean and standard deviation. *[8 marks]*

4. The variance of two data items is k. Find an expression in terms of k for the range. *[5 marks]*

Long questions

1. The following is the cumulative frequency diagram for the heights of 30 plants, given in centimetres.



(a) Use the diagram to estimate the median height.

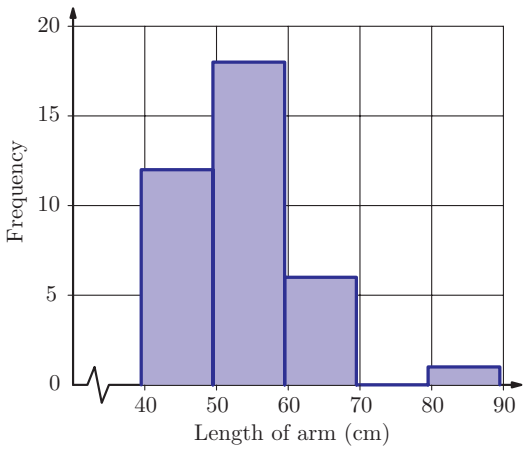
(b) Complete the following frequency table:

Height (h)	Frequency
$0 < h \leq 5$	4
$5 < h \leq 10$	9
$10 < h \leq 15$	
$15 < h \leq 20$	
$20 < h \leq 25$	

(c) Hence estimate the mean height. *[8 marks]*

(© IB Organization 2006)

2. The following histogram shows the length of the arms of 37 children in a classroom in Lithuania, given to the nearest cm.



(a) Explain why the bar representing the first group goes below 40.

(b) Complete the following frequency distribution:

Length	Frequency
40–49	12
50–59	
60–69	
70–79	
80–89	

(c) Use this data to estimate the mean and the standard deviation of the data.

(d) Give one reason to explain why the average arm length of all children in Lithuania might be different from the value found above.

3. The frequency distribution of the number of cars in households on a street is given below:

Number of cars	Frequency
1	a
2	b

(a) Find an expression for the mean of the number of cars and show that the variance is given by $\frac{ab}{(a+b)^2}$.

(b) Prove that it is impossible for the mean to equal the variance.

(c) If the number of households with one car is three times larger than the number of households with two cars find the mean and the standard deviation in the number of cars.

continued . . .

Use the fact that the probabilities add up to 1

$0.7k + 1.2k + 1.44k + 0.84k = 1$
$\therefore k = 0.239$

We can now calculate all the probabilities

We are not asked for exact values, so round them to 3SF

x	1	2	3	4
$P(X = x)$	0.167	0.287	0.344	0.201

One of the most obvious questions to ask about a random variable is what value it is most likely to have. This value is called the **mode**. The random variable X in the above example has mode 3; the most likely number of chocolates you will win is three. A random variable may not have a mode (for example, the outcomes of a fair die are all equally likely) or it may have more than one mode. In particular, if the largest probability corresponds to two of the outcomes, the random variable is called **bimodal**.

Another question we could ask is, if we were to play the above game many times, on average how many chocolates would we expect to win? The answer is not necessarily the same as the most likely outcome. We will see how to answer this question in the next section.

Exercise 23A

1. For each of the following, draw out a table to represent the probability distribution of the random variable described:

(a) A fair coin is thrown four times. The random variable W is the number of tails obtained.

(b) Two fair dice are thrown. The random variable D is the difference between the larger and the smaller score, or zero if they are the same.

(c) A fair die is thrown once. The random variable X is calculated as half the result if the die shows an even number, or one higher than the result if the die shows an odd number.

(d) A bag contains six red and three green counters. Two counters are drawn at random from the bag without replacement. G is the number of green counters remaining in the bag.

In this exercise you will need to use ideas from chapter 22, particularly tree diagrams. For Question 2(c) you may want to look at chapter 7 on Geometric sequences.

(e) Karl picks a card at random from a standard pack of 52 cards. If he draws a diamond, he stops; otherwise, he replaces the card and continues to draw cards at random, with replacement, until he has either drawn a diamond or has drawn a total of 4 cards. The random variable C is the total number of cards drawn.

(f) Two fair four-sided spinners, each labelled 1, 2, 3 and 4, are spun. The random variable X is the product of the two values shown.

2. Find the missing value k:

(a) (i)

x	3	7	9	11
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	k

(ii)

x	5	6	7	10
$P(X = x)$	0.2	0.3	k	0.5

(b) (i) $P(Y = y) = ky$ for $y = 1, 2, 3, 4$

(ii) $P(X = x) = \frac{k}{x}$ for $y = 1, 2, 3, 4$

(c) (i) $P(X = x) = k(0.1)^x$ for $x \in \mathbb{N}$

(ii) $P(R = r) = k(0.9)^r$ for $y \in \mathbb{N}$

3. In a game a player rolls a biased four-sided die. The probability of each possible score is shown below.

Score	1	2	3	4
Probability	$\frac{1}{3}$	$\frac{1}{4}$	k	$\frac{1}{5}$

Find the probability that the total score is four after two rolls. *[5 marks]*

23B Expectation, median and variance of a discrete random variable

The **expectation** of a random variable is a value which represents the mean result if the variable were to be repeatedly measured an infinite number of times. It is a representation of the 'average' value of the random variable.

KEY POINT 23.2

The expected value of a discrete random variable X is written $\mathrm{E}(X)$ and calculated as:

$$\mathrm{E}(X) = \sum_x x\mathrm{P}(X = x)$$

EXAM HINT

In the Formula booklet, μ is also used to denote $\mathrm{E}(X)$.

Worked example 23.3

The random variable X has probability distribution as shown in the table below. Calculate $\mathrm{E}(X)$.

x	1	2	3	4	5	6
$\mathrm{P}(X = x)$	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{10}$

Apply the formula

$$\mathrm{E}(X) = 1\times\frac{1}{10} + 2\times\frac{1}{4} + 3\times\frac{1}{10} + 4\times\frac{1}{4} + 5\times\frac{1}{5} + 6\times\frac{1}{10}$$

$$= \frac{7}{2}$$

Just as the mean of a set of integers could be fractional, so the expectation of a random variable need not be a value which the variable can itself take.

To find the median of a discrete random variable we use the defining property of the median – that half of the data should fall below it. In the context of a probability distribution this means that:

Median, m, is the smallest value of X for which $\mathrm{P}(X \leq m)$ is more than $\frac{1}{2}$. If there is a value m such that $\mathrm{P}(X \leq m) = \frac{1}{2}$ then the median is the mean of this value and the next largest value of X.

Probabilities of the form $\mathrm{P}(X \leq x)$ which give the probability of being less than or equal to a certain value are called **cumulative probabilities**.

Worked example 23.4

Find the median of the probability distribution below:

x	1	3	6	8
$P(X = x)$	0.2	0.4	0.3	0.1

To find the median evaluate the probability of being below each value until you get above 0.5

$P(X \leq 1) = 0.2$

$P(X \leq 3) = 0.6$

Therefore the median is 3.

In the above example if the distribution had been

x	1	3	6	8
$P(X = x)$	0.2	0.3	0.4	0.1

then $P(X \leq 3)$ is exactly 0.5. The median is the average of 3 and 6, so it is 4.5.

As well as knowing the expectation and median, we may also be interested in how far away from the average we can expect an outcome to be. The variance of a random variable is a value representing the degree of variation that would be seen if the variable were to be repeatedly measured an infinite number of times. It is related to how spread out the variable is.

Standard deviation is a much more meaningful representation of the spread of the variable. So why do we bother with variance at all? The answer is purely to do with mathematical elegance. If you do the statistics option (Topic option 7) you will see that the algebra of variance is far neater than the algebra of standard deviations.

This is the same idea as the variance of the set of data from Section 21B.

KEY POINT 23.3

The variance of a random variable X is written Var (X) and is calculated as

$$\text{Var}(X) = \text{E}(X^2) - \left[\text{E}(X)\right]^2$$

where $\text{E}(X^2) = \sum x^2 \text{P}(X = x)$

This formula is often quoted as 'the mean of the squares minus the square of the mean'.

The formula booklet also shows the alternative formula, $\text{E}(X - \mu)^2$, but this is hardly ever used.

Worked example 23.5

Calculate Var(X) for the probability distribution in Worked example 23.3.

Find the expectation

From above, $E(X) = 3.5$

Apply the values from the distribution

$$E(X^2) = 1^2 \times \frac{1}{10} + 2^2 \times \frac{1}{4} + 3^2 \times \frac{1}{10} + 4^2 \times \frac{1}{4} + 5^2 \times \frac{1}{5} + 6^2 \times \frac{1}{10}$$
$$= 14.6$$

$$Var(X) = E(X^2) - [E(X)]^2$$
$$= 14.6 - 12.25 = 2.35$$

Exercise 23B

1. Calculate the expectation, median and variance of each of the following random variables:

(a) (i)

x	1	2	3	4
$P(X = x)$	0.4	0.3	0.2	0.1

(ii)

w	8	9	10	11
$P(W = w)$	0.4	0.3	0.2	0.1

(b) (i) $P(X = x) = \frac{x^2}{14}$, $x = 1, 2, 3$

(ii) $P(X = x) = \frac{1}{x}$, $x = 2, 3, 6$

2. A discrete random variable X is given by

$$P(X = x) = k(x + 1) \text{ for } x = 2, 3, 4, 5, 6.$$

(a) Show that $k = 0.04$.

(b) Find $E(X)$. *[5 marks]*

3. The discrete random variable V has the probability distribution shown below and $E(V) = 6.1$. Find the value of k and the median of V.

v	1	2	5	8	k
$P(V = v)$	0.2	0.3	0.1	0.1	0.3

[6 marks]

4. A discrete random variable X has its probability mass function given by

$$P(X = x) = k(x + 3), \text{ where } x \text{ is } 0, 1, 2, 3.$$

(a) Show that $k = \frac{1}{18}$.

(b) Find the exact value of $E(X)$. *[6 marks]*

5. The probability distribution of a discrete random variable X is defined by:

$$P(X = x) = kx(4 - x), x = 1, 2, 3$$

(a) Find the value of x.

(b) Find $E(X)$. *[6 marks]*

6. A fair six-sided die, with sides numbered 1, 1, 2, 2, 2, 5 is thrown. Find the mean and variance of the score. *[6 marks]*

7. The table below shows the probability distribution of a discrete random variable X.

x	0	1	2	3
$P(X = x)$	0.1	p	q	0.2

(a) Given that $E(X) = 1.5$, find the values of p and q.

(b) Calculate $\text{Var}(X)$. *[9 marks]*

8. A biased die with four faces is used in a game. A player pays 5 counters to roll the die. The table below shows the possible scores on the die, the probability of each score and the number of counters the player wins for each score.

Score	1	2	3	4
Probability	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{20}$
Number of counters player receives	4	5	15	n

Find the value of n in order for the player to get an expected return of 3.25 counters per roll. *[5 marks]*

9. In a game a player pays an entrance fee of $\$n$. He then selects one number from 1, 2, 3 or 4 and rolls three standard dice.

If his chosen number appears on all three dice he wins four times his entrance fee.

If his number appears on exactly two of the dice he wins three times the entrance fee.

If his number appears on exactly one die he wins \$1.

If his number does not appear on any of the dice he wins nothing.

(a) Copy and complete the probability table below.

Profit (\$)	$-n$		$2n$	$3n$
Probability		$\frac{27}{64}$		

(b) The game organiser wants to make a profit over many plays of the game. Given that he must charge a whole number of cents, what is the minimum amount the organiser must charge? *[10 marks]*

23C The binomial distribution

Some discrete probability distributions are met so often that they have been given names and formal notation. One of the most important of these is the **binomial distribution**. There are several others, some of which you will meet in this chapter and some if you study the statistics option (Topic option 7).

A binomial distribution occurs in situations where you have a set number of 'experiments' (or 'trials') each of which have two possible outcomes. The number of trials is usually denoted n. One outcome is conventionally called a 'success' and the other a 'failure'. The probability of success is denoted p. If the probability of success in a trial is constant, and trials are conducted independently of each other, then the number of successes can be modelled using the binomial distribution.

The symbol ~ is used to denote the concept 'follows this distribution', and one or two letter abbreviations are used for the standard distributions. So if a random variable X follows the binomial distribution with n trials and probability of success p, we would write $X \sim \mathrm{B}(n, p)$.

So what is this distribution? Let us consider a specific example: suppose a die is rolled four times, what is the probability of getting exactly two fives? There are four trials so $n = 4$ and if we label a five as a success then $p = \frac{1}{6}$. The probability of a failure is therefore $\frac{5}{6}$.

One way of getting two fives is if the first two times we get a five and the last two times we get something else. The probability of this happening is $\frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{5}{6}$. But this is not the only way

Counting the number of possible selections was discussed in chapter 1.

in which two fives can occur. The two fives may be first and third or second and fourth. In fact, we have to consider all the ways in which we pick two trials out of the four for the 5 to occur. This can happen in $\binom{4}{2}$ ways. Each of them has the same probability as the first case. If X is the random variable 'number of 5s thrown when four dice are rolled' then we can say that:

$$P(X=2)=\binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2$$

The useful thing about identifying a binomial distribution is that you can then apply standard results without having to go through this argument every time. In particular, the expectation and variance of the binomial distribution can just be quoted using the formulae below. The proofs of these are beyond what is expected in the International Baccalaureate®, but if you are interested they are on Fill-in proof 25 'Expectation and variance of the binomial distribution' on the CD-ROM.

KEY POINT 23.4

Standard results of the binomial distribution

Statement of distribution	$X \sim B(n, p)$
Probability formula	$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$
Expectation (E(X))	np
Variance (Var(X))	$np(1-p)$

(Note: in the Formula booklet, the expectaion is referred to as the mean)

Worked example 23.6

Rohir has a 30% chance of correctly answering a multiple-choice question. In a test there are ten questions.

(a) What is the probability that Rohir gets exactly four of them correct? Give your answer to five significant figures.

(b) What is the probability that Rohir gets at least one correct in the first five questions?

(c) Suggest which requirements for a binomial distribution might not be satisfied in this situation?

continued . . .

Define the random variable if not already defined in the question

(a) Let X be the number of correct answers in the first ten questions

Give the probability distribution, checking that the conditions are met

$X \sim B(10, 0.3)$

Express the formula for the probability required, and calculate the answer

$$P(X=4) = \binom{10}{4}(0.3)^4(0.7)^6 = 0.20012 \text{ (5SF)}$$

Define the random variable if not already defined in the question

(b) Let Y be the number of correct answers in the first five questions

Give the probability distribution

$Y \sim B(5, 0.3)$

Write down the probability required

We are interested in $X \geq 1$, which means that $X = 1, 2, 3, 4$ or 5. Remember that a quicker way to do the calculation is to find $1 - P(X < 1)$

$$P(X \geq 1) = 1 - P(X = 0)$$

Express the formula for the probability required, and calculate the answer

$$= 1 - \binom{5}{0}0.3^0\,0.7^5$$

$$= 0.832 \text{ (3SF)}$$

Consider the requirements for the distribution

(c) Binomial requires:

- two outcomes at each trial
- constant probability of success in each trial
- trial results independent of each other

Identify a requirement which is failed in this context: there are two outcomes, and trials are independent (answering one question does not make it easier or harder to answer another)

All questions are not of the same difficulty, so there might not be a constant probability of success.

If you need to find a probability of a range of successes, you could in theory add up the probabilities of individual outcomes. This can be very time consuming, so your calculator has a function giving the probability of getting up to and including any number of successes.

EXAM HINT

Most GDCs can calculate binomial probabilities automatically given n and p, see Calculator sheet 13 on the CD-ROM. But you may also be tested on applying the formula, which is given in the Formula booklet.

Worked example 23.7

Random variable X has distribution B(15, 0.6). Find $P(5 < X \leq 10)$.

The calculator can give us probabilities of the form $P(X \leq k)$

$X \sim B(15, 0.6)$

$P(5 < X \leq 10) = P(X \leq 10) - P(X \leq 5)$

$= 0.7827 - 0.0338$

$= 0.749$ (3SF) (from GDC)

EXAM HINT

Even when you are using a calculator to find probabilities, you should still use correct mathematical notation (not calculator notation) in your answer. You do not need to explain how you did things on the calculator – just state the distribution you used, the probabilities calculated, and give the answer (usually to 3 significant figures).

Exercise 23C

Remember to round your answer to three significant figures when using the calculator.

1. The random variable X has a binomial distribution with $n = 8$ and $p = 0.2$. Calculate:

(a) (i) $P(X = 3)$ (ii) $P(X = 4)$

(b) (i) $P(X \leq 3)$ (ii) $P(X \leq 2)$

(c) (i) $P(X > 3)$ (ii) $P(X > 4)$

(d) (i) $P(X < 5)$ (ii) $P(X < 3)$

(e) (i) $P(X \geq 3)$ (ii) $P(X \geq 1)$

(f) (i) $P(3 < X \leq 6)$ (ii) $P(1 \leq X < 4)$

2. Given that $Y \sim B(5,\ 0.5)$, find the exact value of:

(a) (i) $P(Y = 1)$ (ii) $P(Y = 0)$

(b) (i) $P(Y \geq 1)$ (ii) $P(Y \leq 1)$

(c) (i) $P(Y > 4)$ (ii) $P(Y \leq 3)$

3. Find the mean and standard deviation of each of the following variables:

(a) (i) $Y \sim B\left(100,\ \frac{1}{10}\right)$ (ii) $X \sim B\left(16,\ \frac{1}{2}\right)$

(b) (i) $X \sim B(15,\ 0.3)$ (ii) $Y \sim B(20,\ 0.35)$

(c) (i) $Z \sim B\left(n-1,\ \frac{1}{n}\right)$ (ii) $X \sim B\left(n,\ \frac{2}{n}\right)$

4. (a) Jake beats Marco at chess in 70% of their games. Assuming that this probability is constant and that the results of games are independent of each other, what is the probability that Jake will beat Marco in at least 16 of their next 20 games?

(b) On a television channel, the news is shown at the same time each day; the probability that Salia watches the news on a given day is 0.35. Calculate the probability that on 5 consecutive days she watches the news on exactly 3 days.

(c) Sandy is playing a computer game and needs to accomplish a difficult task at least three times in five attempts in order to pass the level. There is a 1 in 2 chance that he accomplishes the task each time he tries, unaffected by how he has done before. What is the probability that he will pass to the next level?

5. 15% of students at a large school travel by bus. A random sample of 20 students is taken.

(a) Explain why the number of students in the sample who travel by bus is only approximately a binomial distribution.

(b) Use the binomial distribution to estimate the probability that exactly five of the students travel by bus. *[3 marks]*

6. A coin is biased so that when it is tossed the probability of obtaining heads is $\frac{2}{3}$. The coin is tossed 4050 times. Let X be the number of heads obtained. Find:

(a) the mean of X

(b) the standard deviation of X. *[3 marks]*

7. A biology test consists of eight multiple-choice questions. Each question has four answers, only one of which is correct. At least five correct answers are required to pass the test. Sheila does not know the answers to any of the questions, so answers each question at random.

(a) What is the probability that Sheila answers exactly five questions correctly?

(b) What is the expected number of correct answers Sheila will give?

(c) What is the standard deviation in the number of correct answers Sheila will give?

(d) What is the probability that Sheila manages to pass the test? *[7 marks]*

8. 0.8% of people in the country have a particular cold virus at any time. On a single day, a doctor sees 80 patients.

(a) What is the probability that exactly 2 of them have the virus?

(b) What is the probability that 3 or more of them have the virus?

(c) State an assumption you have made in these calculations. *[5 marks]*

9. Given that $Y \sim B(12,0.4)$:

(a) Find the expected mean of Y.

(b) Find the mode of Y. *[3 marks]*

10. On a fair die, which is more likely: rolling 3 sixes in 4 throws or rolling a five or a six in 5 out of 6 throws? *[6 marks]*

11. Over a one month period, Ava and Sven play a total of x games of tennis. The probability that Ava wins any game is 0.4. The result of each game played is independent of any other game played. Let X denote the number of games won by Ava over a one month period.

(a) Find an expression for $P(X = 2)$ in terms of n.

(b) If the probability that Ava wins two games is 0.121 correct to three decimal places, find the value of n.

[5 marks]

12. A coin is biased so that the probability of it showing tails is p. The coin is tossed n times. Let X be a random variable representing the number of tails. It is known that the mean of X is 19.5 and the variance is 6.825. Find the values of n and p. *[5 marks]*

13. A die is biased so that the probability of rolling a 6 is p. If the probability of rolling 2 sixes in 12 throws is 0.283 (to three significant figures), find the possible values of p correct to two decimal places. *[5 marks]*

14. In an experiment, a trial is repeated n times. The trials are independent and the probability p of success in each trial is constant. Let X be the number of successes in the n trials. The mean of X is 12 and the standard deviation is 2. Find n and p. *[5 marks]*

15. X is a binomial random variable, where the number of trials is 4 and the probability of success of each trial is p. Find the possible values of p if $P(X = 3) = 0.3087$. *[5 marks]*

16. X is a binomial random variable, where the number of trials is 4 and the probability of success of each trial is p. Find the possible values of p if $P(X = 2) = \frac{96}{625}$. *[6 marks]*

Question 10 is the problem which was posed to Pierre de Fermat in 1654 by a professional gambler who could not understand why he was losing. It inspired Fermat (with the assistance of Pascal) to set up probability as a rigorous mathematical discipline.

23D The Poisson distribution

When you are waiting for a bus there are at any given moment two possible outcomes – it either arrives or it does not. We can try modelling this situation using a binomial distribution, but it is not clear what an individual trial is. Instead we have a rate of success – the number of buses that arrive in a fixed time period.

EXAM HINT

If a question mentions average rate of success, or events occurring at a constant rate, you should use Poisson distribution. If you can identify a fixed number of trials then binomial distribution is required.

There are many situations in which we know the rate of events within a given space or time, in contexts ranging from commercial (such as the number of calls through a telephone exchange per minute) to biological (such as the number of clover plants seen per square metre in a pasture). Where the events occur singly (one at a time) and can be considered independent of each other (so that the probability of each event is not affected by what has already happened), the number of events in a fixed space or time interval can be modelled using **Poisson distribution**. This distribution is fully defined once we know the rate of success, which is conventionally called m.

KEY POINT 23.5

Standard results of the Poisson distribution

Statement of distribution	$X \sim \text{Po}(m)$
Probability formula	$\text{P}(X = x) = \dfrac{\text{e}^{-m}m^x}{x!}$ *for* $x = 0,1,2, \ldots$
Expectation E(X)	m
Variance Var (X)	m

(Note: in the Formula booklet, E(X) is called the mean)

Worked example 23.8

Recordable accidents occur in a factory at an average rate of 7 every year, independently of each other. Find the probability that in a given year exactly 3 recordable accidents occurred.

Define the random variable

Let X be the number of accidents in a year

Give the probability distribution

$X \sim \text{Po}(7)$

Write down the probability required, and calculate the answer

$$\text{P}(X = 3) = \frac{e^{-7}7^3}{3!}$$

$$= 0.521 \text{ (3SF)}$$

The Poisson distribution is scaleable. If the number of butterflies seen on a flower in 10 minutes follows a Poisson distribution with mean (expectation) m, then the number of butterflies seen on a flower in 20 minutes follows a Poisson distribution with mean $2m$; the number of butterflies seen on a flower in 5 minutes follows a Poisson distribution with mean $\frac{m}{2}$.

EXAM HINT

See Calculator sheet 13 on the CD-ROM. Your GDC can calculate Poisson probabilities and cumulative probabilities, but you may be explicitly asked to use the formula. Remember to round your answers to 3SF.

Worked example 23.9

If there are an average of 12 buses per hour arriving at a bus stop, find the probability that there are more than 6 buses in 30 minutes.

Define the random variable

Let X be the number of buses in 30 minutes

Give the probability distribution

$X \sim Po(6)$

Write down the probability required. To use the calculator we must relate it to $P(X \leq k)$

$P(X > 6) = 1 - P(X \leq 6)$

$= 0.161$ (3SF) from GDC

Exercise 23D

1. State the distribution of the variable in each of the following situations:

(a) Cars pass under a motorway bridge at an average rate of 6 per 10 second period.

(i) the number of cars passing under the bridge in 1 minute

(ii) the number of cars passing under the bridge in 15 seconds

(b) Leaks occur in water pipes at an average rate of 12 per kilometre.

(i) the number of leaks in 200 m

(ii) the number of leaks in 10 km

(c) A widget machine manufactures on average 96 functional widgets out of 100.

(i) the number of faulty widgets in a sample of 10

(ii) the number of functioning widgets in sample of 20

(d) 12 worms are found on average in a 1 m^2 area of a garden.

(i) the number of worms found in a 0.3 m^2 area

(ii) the number of worms found in a 2 m by 2 m area

2. Calculate the following probabilities:

(a) If $X \sim \text{Po}(2)$

(i) $\text{P}(X = 3)$ (ii) $\text{P}(X = 1)$

(b) If $Y \sim \text{Po}(1.4)$

(i) $\text{P}(Y \leq 3)$ (ii) $\text{P}(Y \leq 1)$

(c) If $Z \sim \text{Po}(7.9)$

(i) $\text{P}(Z < 6)$ (ii) $\text{P}(Z < 10)$

(d) If $X \sim \text{Po}(5.9)$

(i) $\text{P}(X \geq 3)$ (ii) $\text{P}(X > 1)$

(e) If $X \sim \text{Po}(11.4)$

(i) $\text{P}(8 < X < 11)$ (ii) $\text{P}(8 \leq X \leq 12)$

3. A random variable X follows a Poisson distribution with mean 1.7. Copy and complete the following table of probabilities, giving results to 3 significant figures:

x	0	1	2	3	4	> 4
$\text{P}(X = x)$	0.183					

4. From a particular observatory, shooting stars are observed in the night sky at an average rate of one every 5 minutes. Assuming that this rate is constant and that shooting stars occur (and are observed) independently of each other, what is the probability that more than 20 are seen over a period of 1 hour? *[4 marks]*

5. When examining blood from a healthy individual under a microscope, a haematologist knows that on average he should see 4 white blood cells in each high power field. Find the probability that blood from a healthy individual will show:

(a) 7 white blood cells in a single high power field

(b) a total of 28 white blood cells in 6 high power fields, selected independently. *[5 marks]*

6. Salah is sowing flower seeds in his garden. He scatters seeds randomly so that the number of seeds falling on any particular region is a random variable with a Poisson distribution, with mean value proportional to the area. He will sow fifty thousand seeds over an area of 2 m^2.

(a) Calculate the expected number of seeds falling on a 1 cm^2 region.

(b) Calculate the probability that a given 1 cm^2 area receives no seeds. *[4 marks]*

7. A wire manufacturer is looking for flaws. Experience suggests that there are on average 1.8 flaws per metre in the wire.

(a) Determine the probability that there is exactly 1 flaw in 1 metre of the wire.

(b) Determine the probability that there is at least one flaw in 2 metres of the wire. *[5 marks]*

8. The random variable X has a Poisson distribution with mean 5. Calculate:

(a) $P(X \leq 5)$

(b) $P(3 < X \leq 5)$

(c) $P(X \neq 4)$

(d) $P(3 < X \leq 5 \mid X \leq 5)$ *[8 marks]*

9. Patients arrive at random at an emergency room in a hospital at the rate of 14 per hour throughout the day.

(a) Find the probability that exactly 4 patients will arrive at the emergency room between 18:00 and 18:15.

(b) Given that fewer than 15 patients arrive in one hour, find the probability that more than 12 arrive. *[6 marks]*

10. The number of eagles observed in a forest in one day follows a Poisson distribution with mean 1.4.

(a) Find the probability that more than three eagles will be observed on a given day.

(b) Given that at least one eagle is observed on a particular day, find the probability that exactly two eagles are seen that day. *[6 marks]*

11. The random variable X follows a Poisson distribution. Given that $P(X \geq 1) = 0.4$, find:

(a) the mean of the distribution

(b) $P(X > 2)$. *[5 marks]*

12. The random variable X is Poisson distributed with mean m and satisfies $P(X = 3) = P(X < 3)$.

(a) Find the value of m, correct to four decimal places.

(b) For this value of m evaluate $P(2 \leq X < 4)$. *[6 marks]*

13. Let X be a random variable with a Poisson distribution, such that $P(X > 2) = 0.3$. Find $P(X < 2)$. *[5 marks]*

14. The number of emails you receive per day follows a Poisson distribution with mean 6. Let D be the number of emails received in one day and W the number of emails received in a week.

(a) Calculate $P(D = 6)$ and $P(W = 42)$.

(b) Find the probability that you receive 6 emails every day in a seven-day week.

(c) Explain why this is not the same as $P(W = 42)$. *[8 marks]*

15. The number of mistakes a teacher makes whilst marking homework has a Poisson distribution with a mean of 1.6 errors per piece of homework.

(a) Find the probability that there are at least two marking errors in a randomly chosen piece of homework.

(b) Find the most likely number of marking errors occurring in a piece of homework. Justify your answer.

(c) Find the probability that in a class of 12 pupils fewer than half of them have errors in their marking. *[9 marks]*

16. A car company has two limousines that it hires out by the day. The number of requests per day has a Poisson distribution with mean 1.3 requests per day.

(a) Find the probability that neither limousine is hired.

(b) Find the probability that some requests have to be denied.

(c) If each limousine is to be equally used, on how many days in a period of 365 days would you expect a particular limousine to be in use? *[8 marks]*

17. A shop has 4 copies of the book 'Ballroom Dancing' delivered each week. The demand for the book follows a Poisson distribution with mean 3.2 requests per week.

(a) Calculate the probability that the shop cannot meet the demand in a given week.

(b) Find the most probable number of books sold in one week.

(c) Find the expected number of books sold in one week.

(d) Determine the smallest number of copies of the book that should be ordered each week to ensure that the demand is met with a probability of at least 98%. *[8 marks]*

18. The random variable X follows Poisson distribution with mean λ. If $\mathrm{P}(X=2)=\mathrm{P}(X=0)+\mathrm{P}(X=1)$, find the exact value of λ. *[4 marks]*

19. The random variable X follows a Poisson distribution with mean λ.

(a) Show that $\mathrm{P}(Y=y+2)=\dfrac{\lambda^2}{(y+1)(y+2)}\mathrm{P}(Y=y)$.

(b) Given that $\lambda=6\sqrt{2}$, find the value of y such that $\mathrm{P}(Y=y+2)=\mathrm{P}(Y=y)$. *[4 marks]*

Summary

- A **random variable** is a quantity whose value depends on chance. A list of all possible outcomes and their associated probabilities is called a **probability distribution** or **probability mass function**.
- The total of all the probabilities of a probability distribution must always equal 1.
- Even though the outcome of any one observation of a random variable is impossible to predict with any certainty, the **expectation** (of the mean) and variance of observations can be predicted quite accurately, using:

$$\mathrm{E}(X)=\sum_x x\mathrm{P}(X=x)$$

$$\mathrm{Var}(X)=\mathrm{E}(X^2)-\left[\mathrm{E}(X)\right]^2$$

- If there is a fixed number of trials (each with two possible outcomes) with constant and independent probability of success in each trial then the number of successes follows a **Binomial distribution**: $X\sim\mathrm{B}(n,p)$.
- If events occur singly, independently and at a constant rate, then the number of events in a given period follows a **Poisson distribution**: $X\sim\mathrm{Po}(m)$, where m is the **rate of success**.
- Once the distribution has been identified then probabilities and statistics for the distribution can be immediately quoted:

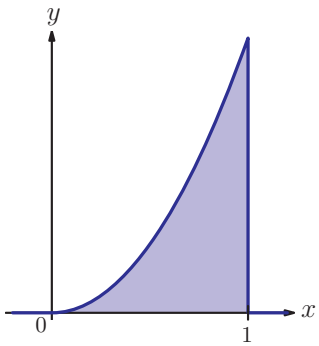
Distribution	Notation	$\mathrm{P}(X=x)$	$\mathrm{E}(X)$	$\mathrm{Var}(X)$
Binomial	$X\sim\mathrm{B}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$
Poisson	$X\sim\mathrm{Po}(m)$	$\dfrac{\mathrm{e}^{-m}m^x}{x!}$	m	m

Worked example 24.1

A continuous random variable has a pdf:

$$f(x)=\begin{cases} kx^2 & 0\le x<1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of k.
(b) Find the probability of x being between 0.2 and 0.6.



Total area is 1. Area is only found between 0 and 1

(a) $1=\int_0^1 kx^2\,dx$

$=\left[\frac{kx^3}{3}\right]_0^1$

$=\frac{k}{3}$

$\Leftrightarrow k=3$

Probability X lies in [a, b] is $\int_a^b f(x)\,dx$

(b) $P(0.2<X<0.6)=\int_{0.2}^{0.6} 3x^2\,dx$

$=[x^3]_{0.2}^{0.6}$

$=0.208$

Exercise 24A

1. For each of these distributions, find the possible values of the unknown parameter k:

(a) (i) $f(x)=kx^3,\ 2<x<3$ (ii) $f(x)=k\sqrt{x},\ 1<x<4$

(b) (i) $f(x)=x^2+k,\ -1<x\le 2$ (ii) $f(x)=3x+k,\ -2\le x<3$

(c) (i) $f(x)=e^{kx},\ 0<x<2$ (ii) $f(x)=\sin kx,\ 0<x<\pi$

(d) (i) $f(x)=\frac{1}{(x+k)^2},\ 0\le x\le 1$ (ii) $f(x)=\frac{1}{x+k},\ 0\le x\le 1$

(e) (i) $f(x)=x^3,\ 0<x<k$ (ii) $f(x)=2x-1,\ 1<x<k$

(f) (i) $f(x)=\frac{1}{1+x},\ k<x<k+1$ (ii) $f(x)=x^2,\ k<x<2k$

(g) (i) $f(x)=kx^2,\ 0<x<k$ (ii) $f(x)=x+k,\ 0<x<k$

(h) (i) $f(x) = ke^{-x^2},\ 3 < x < 8$ (ii) $f(x) = k\sin\sqrt{x},\ \pi < x < \pi^2$

(i) (i) $f(x) = \dfrac{1}{x^2},\ 1 < x < k$ (ii) $f(x) = \dfrac{1}{2\sqrt{x}},\ k < x < 1$

2. (a) If $f(x) = \begin{cases} 2-2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

(i) Find $P(0.3 < X < 0.9)$. (ii) Find $P(0 < X < 0.5)$.

(b) If $f(x) = \begin{cases} \cos x & 0 < x < \dfrac{\pi}{2} \\ 0 & \text{otherwise} \end{cases}$

(i) Find $P\left(\dfrac{\pi}{4} < X \leq \dfrac{\pi}{3}\right)$. (ii) Find $P\left(0 \leq X < \dfrac{\pi}{6}\right)$.

(c) If $f(x) = \begin{cases} \dfrac{1}{x\ln 10} & 1 < x < 10 \\ 0 & \text{otherwise} \end{cases}$

(i) Find $P(X > 5)$. (ii) Find $P(X \leq 3)$.

3. (a) If $f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

(i) Find a if $P(X < a) = 0.4$. (ii) Find b if $P(X < b) = 0.9$.

(b) If $f(x) = \begin{cases} \dfrac{x}{8} & 0 < x < 8 \\ 0 & \text{otherwise} \end{cases}$

(i) Find a if $P(X > a) = 0.9$. (ii) Find b if $P(X > b) = 0.5$.

(c) If $f(x) = \begin{cases} \dfrac{x}{16} & 2 < x < 6 \\ 0 & \text{otherwise} \end{cases}$

(i) Find a if $P(2 + a < X < 6 - a) = 0.8$.
(ii) Find b if $(b < X < b + 1) = 0.25$.

4. A model predicts that the angle, G, by which an alpha particle is deflected by a nucleus is modelled by:

$$f(g) = \begin{cases} kg^2 & 0 < g < \pi \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of the constant k.

(b) 10 000 alpha particles are fired at a nucleus. If the model is correct, estimate the number of alpha particles deflected by less than $\dfrac{\pi}{3}$. *[6 marks]*

5. A random variable Y has distribution:

$$f(y)=\begin{cases}3e^{-3x} & y>0\\ 0 & \text{otherwise}\end{cases}$$

Find the exact value of $P(Y>2)$. *[4 marks]*

6. If the continuous random variable X has a probability density

$$f(x)=\begin{cases}\sec^2 x & 0<x<\dfrac{\pi}{4}\\ 0 & \text{otherwise}\end{cases}$$

find the interquartile range of X. *[6 marks]*

7. If $f(x)=\begin{cases}\dfrac{1}{x} & 1<x<e\\ 0 & \text{otherwise}\end{cases}$

(a) Find b in terms of k if $P(b<X<b^2)=k$.

(b) Find a in terms of k if $P(2-a<X\leq 2+a)=k$. *[7 marks]*

8. If $f(x)=\begin{cases}e^x & k<x<2k\\ 0 & \text{otherwise}\end{cases}$

find $P\left(X>\dfrac{3k}{2}\right)$. *[7 marks]*

24B Expectation and variance of continuous random variables

The expressions for expectation and variance of continuous random variables all involve integration.

KEY POINT 24.3

Expectation and variance of continuous random variables:

$$E(X)=\int_{-\infty}^{\infty} xf(x)\,dx$$

$$E(X^2)=\int_{-\infty}^{\infty} x^2 f(x)\,dx$$

$$\text{Var}(X)=E(X^2)-[E(X)]^2$$

Note: The formulae in the Formula booklet are set out slightly differently.

You may have noticed that the expressions for $E(X)$ and $\text{Var}(X)$ look similar to those for discrete random variables, but with integration signs instead of summation signs. This is because there is a link between sums and integrals, which you met in chapter 17. You will explore this further if you study chapter 28 in the Calculus option (Topic option 9).

Worked example 24.2

If a continuous random variable has pdf

$$f(x)=\begin{cases}\frac{3}{4}x(2-x) & 0<x<2,\\ 0 & \text{otherwise}\end{cases}$$

find E(X) and the standard deviation of X.

We can do the integration on the calculator. If you need reminding how, See Calcultor skills sheet 10 on the CD-ROM

$$E(X)=\int_0^2 x\times\frac{3}{4}x(2-x)\,dx$$
$$=\frac{3}{4}\int_0^2 x^2(2-x)\,dx$$
$$=1 \text{ (from GDC)}$$

To find standard deviation we must first find Var(X) which requires us to find E(X^2)

$$E(X^2)=\int_0^2 x^2\times\frac{3}{4}x(2-x)\,dx$$
$$=\frac{3}{4}\int_0^2 x^3(2-x)\,dx$$
$$=1.2 \text{ (from GDC)}$$
$$\therefore \text{Var}(X)=E(X^2)-E(X)^2$$
$$=1.2-1^2$$
$$=0.2$$
$$\text{standard deviation}=\sqrt{0.2}=0.447$$

EXAM HINT

The expected mean appears in examination questions more often than the median or mode.

It is also possible to find the median and mode for a continuous distribution.

The defining feature of the median is that half of the data should be below this value and half above. The mode is the most likely value. We can interpret this in terms of probability.

KEY POINT 24.4

The median, m, satisfies

$$\int_{-\infty}^{m} f(x)\,dx=\frac{1}{2}$$

The mode is the value of x at the maximum value of $f(x)$.

The maximum value of $f(x)$ is not necessarily where $\frac{df}{dx}=0$, see Section 16H.

Worked example 24.3

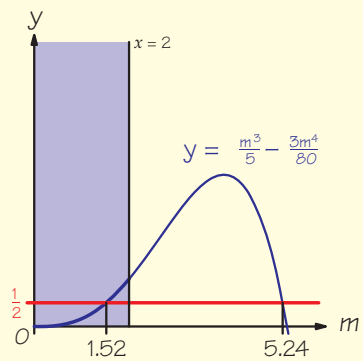
If $f(x) = \begin{cases} \frac{3}{20}(4x^2 - x^3) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$ find the median and mode of X.

Probability of being below the median is $\frac{1}{2}$

$$\int_0^m \frac{3}{20}(4x^2 - x^3)\,dx = \frac{1}{2}$$

$$\Leftrightarrow \frac{m^3}{5} - \frac{3m^4}{80} = \frac{1}{2}$$

This is a quartic equation without any easy substitution. Time to use the calculator



From GDC: m = 1.52 or 5.24

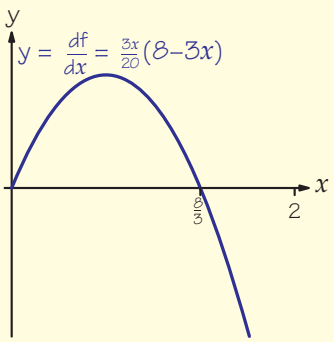
However $0 < m < 2$ therefore median = 1.52

For the mode check for a local maximum

$$\frac{df}{dx} = \frac{6x}{5} - \frac{9x^2}{20} = 0$$

$$= \frac{3x}{20}(8 - 3x)$$

$$\Leftrightarrow x = 0 \text{ or } x = \frac{8}{3}$$



From the graph of f(x) it is clear that $x = \frac{8}{3}$ corresponds to a maximum, so the mode is $\frac{8}{3}$

Exercise 24B

1. Find E(X), the median of X, the mode of X and Var(X) if X has the given probability density function:

(a) (i) $f(x) = 2 - 2x \quad 0 < x < 1$ (ii) $f(x) = \frac{x}{8} \quad 0 < x < 8$

(b) (i) $f(x) = \frac{1}{x \ln 10} \quad 1 < x < 10$ (ii) $f(x) = \frac{2}{x^2} \quad 1 < x < 2$

(c) (i) $f(x) = \cos x \quad 0 < x < \frac{\pi}{2}$ (ii) $f(x) = e^x \quad 0 < x < \ln 2$

(d) (i) $f(x) = \frac{3}{x^4} \quad x > 1$ (ii) $f(x) = \frac{4}{x^5} \quad x > 1$

2. (a) Given that E(X) = 1.1, find k if:

(i) $f(x) = \begin{cases} \frac{1}{x \ln k} & 1 < x < k \\ 0 & \text{otherwise} \end{cases}$ (ii) $f(x) = \begin{cases} \frac{k}{x^k} & 1 < x < \infty \\ 0 & \text{otherwise} \end{cases}$

(b) Given that E(X) = 3, find k if:

(i) $f(x) = \begin{cases} k & k < x < k + \frac{1}{k} \\ 0 & \text{otherwise} \end{cases}$

(ii) $f(x) = \begin{cases} \frac{1}{x + k} & 0 < x < (e - 1) \\ 0 & \text{otherwise} \end{cases}$

3. The continuous random variable X has pdf:

$$f(x) = \begin{cases} \frac{3}{20}(4x^2 - x^3) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the expected mean of X.

(b) Find the mode of X. *[6 marks]*

4. A continuous random variable B has pdf:

$$f(b) = \begin{cases} ab^2 & 3 < b < 10 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of the constant a.

(b) Find E(B). *[7 marks]*

5. A function f is given by:

$$f(y) = \begin{cases} ke^{-kx} & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that f is a probability density function

(b) A random variable Y has distribution given by $f(x)$. Find E(Y) in terms of k. *[10 marks]*

6. Y is a continuous random variable with probability density function:

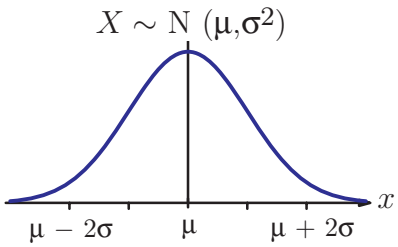
$$f(y)=\begin{cases} ay^2 & -k<y<k \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that $a=\dfrac{3}{2k^3}$.

(b) Given that Var(Y) = 5 find the exact value of k. *[11 marks]*

7. Given that $f(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}, x\in\mathbb{R}$ is a probability density function find E(X) and prove that Var(X) = 1. *[9 marks]*

24C The normal distribution



There are many situations where a variable is most likely to be close to its average value, and values further away from the average become increasingly unlikely. Many such situations can be modelled using the **normal distribution.**

All that is needed to describe this distribution is its mean (μ) and variance (σ^2). If a variable follows this distribution we use the notation $X \sim N(\mu, \sigma^2)$.

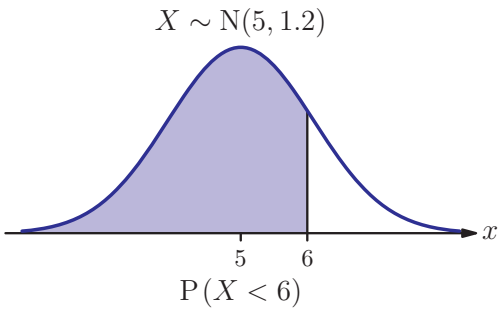
The probability density function (pdf) for the normal distribution is quite complicated:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This function cannot be integrated in terms of other well-known functions, but your calculator can find approximate probabilities.

See Calculator sheet 14 on the CD-ROM.

You may find it helpful to sketch a diagram to get a visual representation of the probability you are trying to find.

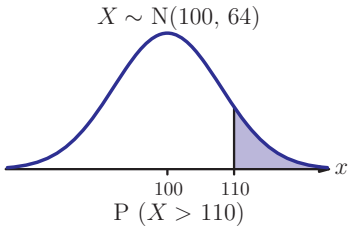


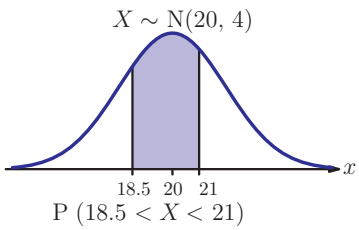
EXAM HINT

Be careful with the notation: σ^2 is the variance, so $X \sim N(10, 9)$ has standard deviation $\sigma = 3$.

Historically, cumulative probabilities for the normal distribution were recorded in tables and these are still used if you don't have a graphical calculator. As there cannot be separate tables for every possible different μ and σ, all values needed to be converted into a Z-score described later.

The diagrams can also provide a useful check, to see whether you should expect the probability to be smaller or greater than 0.5.





Worked example 24.4

The average height of people in a town is 170 cm with standard deviation of 10 cm. What is the probability that a randomly selected resident:

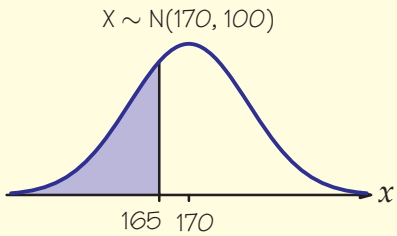
(a) is less than 165 cm tall?
(b) is between 180 cm and 190 cm tall?
(c) is over 176 cm tall?

State the distribution used

X is the crv 'height of a town resident' so
$X \sim N(170, 100)$

State the probability to be found and use the calculator

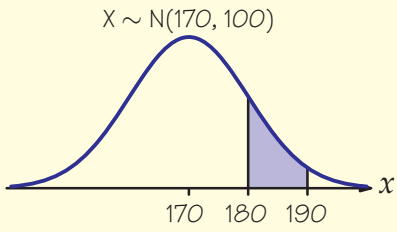
(a) $p(X < 165)$



$P(X < 165) = 0.309 (3SF)$ (from GDC)

State the probability to be found and use the calculator

(b) $p(180 < X < 190)$

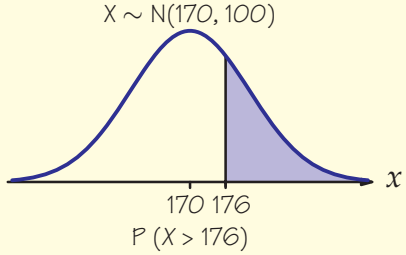


$P(180 < X < 190) = 0.136 (3SF)$ (from GDC)

continued . . .

State the probability to be found and use the calculator

(c) $X \sim N(170, 100)$



$P(X > 176) = 0.274 (3SF)$ (from GDC)

If a normally distributed random variable has mean 120, should a value of 150 be considered unusually large? The answer depends on how spread out the variable is, which is measured by its standard deviation. If the standard deviation is 30 then a value around 150 will be quite common; however, if the standard deviation were 5 then 150 would be very unusual.

It turns out that the probability of a normally distributed random variable being less than a given value ($P(X \leq x)$, called the **cumulative probability**) depends only on the number of standard deviations x is from the mean. This is called the **Z-score**.

In the real world, there is always a possibility that a result may have occurred by random chance. Supplementary sheet 12 'Significant discoveries' explores how unlikely a result has to be before we accept it was not a fluke, which is often stated in terms of the z-score.

KEY POINT 24.5

For $X \sim N(\mu, \sigma^2)$, the Z-score measures the number of standard deviations of x above the mean.

$$z = \frac{x - \mu}{\sigma}$$

(a negative Z-score means x is below the mean)

Worked example 24.5

Given that $X \sim N(15, 6.25)$:

(a) How many standard deviations is $x = 16.1$ away from the mean?
(b) Find the value of X which is 1.2 standard deviations below the mean.

The number of standard deviations away from the mean is measured by the Z-score

(a) $z = \frac{x - \mu}{\sigma}$

continued . . .

6.25 is the variance

$\sigma = \sqrt{6.25} = 2.5$

$\therefore z = \dfrac{16.1 - 15}{2.5} = 0.44$

16.1 is 0.44 standard deviations away from the mean.

Values below the mean have a negative Z-score

(b) $z = -1.2$

$-1.2 = \dfrac{x - 15}{2.5}$

$\Rightarrow x - 15 = -3$

$\Rightarrow x = 12$

Before graphical calculators existed (which wasn't so long ago!) people used tables showing cumulative probabilities of the standard normal distribution. Because of their importance they were given special notation: $\Phi(z) = P(Z \leq z)$. Although you do not have to use this notation, you should understand what it means.

If we are given a random variable $X \sim N(\mu, \sigma^2)$ we can create a new random variable Z which takes the values equal to the Z-scores of the values of X. In other words, for each x there is a corresponding $z = \dfrac{x - \mu}{\sigma}$. This is called the standardised value.

It turns out that, whatever the original mean and standard deviation of X, this new random variable always has normal distribution with mean 0 and variance 1, called the **standard normal distribution**: $Z \sim N(0, 1)$. This is an extremely important property of the normal distribution which needs to be used in situations when the mean and standard deviation of X are not known (see next section).

KEY POINT 24.6

The probabilities of X and Z are related by

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right)$$

Worked example 24.6

Let $X \sim N(6, 0.5^2)$. Write the following in terms of probabilities of Z:

(a) $P(X \leq 6.1)$
(b) $P(5 < X < 7)$
(c) $P(X > 6.5)$

We are given that $x = 6.1$ so we can calculate z

(a) $P(X \leq 6.1) = P\left(Z \leq \dfrac{6.1 - 6}{0.5}\right) = P(Z \leq 0.2)$

continued . . .

The relationship between X and Z above is stated for probabilities of the form $P(X \leq k)$, so convert to that form first

(b) $P(5 < X < 7) = P(X < 7) - P(X < 5)$

$= P\left(Z < \frac{7-6}{0.5}\right) - P\left(Z < \frac{5-6}{0.5}\right)$

$= P(Z < 2) - P(Z < -2) = P(-2 < Z < 2)$

(c) $P(X > 6.5) = 1 - P(X \leq 6.5)$

$= 1 - P\left(Z \leq \frac{6.5-6}{0.5}\right) = 1 - P(Z \leq 1)$

$= P(Z > 1)$

You can see from the examples above that you don't actually have to convert probabilities into the form $P(X \leq k)$ every time; simply replace the x values by the corresponding z scores.

Exercise 24C

1. Find the following probabilities:

(a) If $X \sim N(20,100)$,

(i) $P(X \leq 32)$ (ii) $P(X < 12)$

(b) If $Y \sim N(4.8,1.44)$,

(i) $P(Y > 5.1)$ (ii) $P(Y \geq 3.4)$

(c) If $R \sim N(17,2)$

(i) $P(16 < R < 20)$ (ii) $P(17.4 < R < 18.2)$

(d) If Q has a normal distribution with mean 12 and standard deviation 3:

(i) $P(Q > 9.4)$ (ii) $P(Q < 14)$

(e) If F has a normal distribution with mean 100 and standard deviation 25:

(i) $P(|F - 100| < 15)$ (ii) $P(|F - 100| > 10)$

2. Find the Z-score corresponding to the given value of X:

(a) (i) $X \sim N(12, 2^2), x = 13$ (ii) $X \sim N(38, 7^2), x = 45$

(b) (i) $X \sim N(20, 9), x = 15$ (ii) $X \sim N(162, 25), x = 160$

3. Given that $X \sim N(16,\ 2.5^2)$, write the following in terms of probabilities of the standard normal variable:

(a) (i) $P(X < 20)$ (ii) $P(X < 19.2)$

(b) (i) $P(X \geq 14.3)$ (ii) $P(X \geq 8.6)$

(c) (i) $P(12.5 < X < 16.5)$ (ii) $P(10.1 \leq X \leq 15.5)$

4. It is found that the lifespan of a certain brand of laptop batteries follows normal distribution with mean 16 hours and standard deviation 5 hours. A particular battery has a lifespan of 10.2 hours.

(a) How many standard deviations below the mean is this?

(b) What is the probability that a randomly chosen laptop battery has a lifespan shorter than this? *[6 marks]*

5. When Ali competes in long-jump competitions, the lengths of his jumps are normally distributed with mean 5.2 m and standard deviation 0.7 m.

(a) What is the probability that Ali will record a jump between 5 m and 5.5 m?

(b) Ali needs to jump 6 m to qualify for the school team.

(i) What is the probability that he will qualify with a single jump?

(ii) If he is allowed three jumps, what is the probability that he will qualify for the school team? *[7 marks]*

6. Weights of a species of cat have a normal distribution with mean 16 kg and variance 16 kg^2. In a sample of 2000 such cats, estimate the number which will have a weight above 13 kg. *[6 marks]*

You saw in chapter 22 that ∩ means intersection.

7. If $D \sim N(250, 400)$, find:

(a) $P(D > 265 \cap D < 280)$

(b) $P(D > 265 \mid D < 280)$

(c) $P(D < 242 \cap D > 256)$ *[6 marks]*

8. If $Q \sim N(4,160)$, find:

(a) $P(5 < |Q|)$

(b) $P(Q > 5 \mid 5 < |Q|)$ *[6 marks]*

9. The weights of apples are normally distributed with mean weight 150 g and standard deviation 25 g. Supermarkets classify apples as 'medium' if they are between 120 g and 170 g.

(a) What proportion of apples are medium?

(b) In a bag of 10 apples what is the probability that there are at least 8 medium apples? *[6 marks]*

10. The wingspans of a species of pigeon are normally distributed with mean length 60 cm and standard deviation 6 cm. A pigeon is chosen at random.

(a) Find the probability that its wingspan is greater than 50 cm.

(b) Given that its length is greater than 50 cm, find the probability that a wingspan is greater than 55 cm. *[6 marks]*

11. Grains of sand are believed to have a normal distribution with mean 2 mm and variance 0.25 mm^2.

(a) Find the probability that a randomly chosen grain of sand is larger than 1.5 mm.

(b) The sand is passed through a filter which blocks grains wider than 2.5 mm. The sand that passes through the filter is examined. What is the probability that a randomly chosen grain of filtered sand is larger than 1.5 mm? *[6 marks]*

12. The amount of paracetamol per tablet is believed to be normally distributed with mean 500 mg and standard deviation 160 mg. A dose of less than 300 mg is ineffective in dealing with toothache. In a trial of 20 people treated for toothache with a single tablet, what is the probability that 2 or more of them have less than the effective dose? *[6 marks]*

13. A variable has a normal distribution with a mean that is 7 times its standard deviation. What is the probability of the variable taking a value less than 5 times the standard deviation? *[6 marks]*

14. If $X \sim N(\mu, \sigma^2)$ and $P(X \leq x) = k$ find $P(X \leq 2\mu - x)$ in terms of k. *[5 marks]*

24D Inverse normal distribution

In Section C we saw how to find probabilities when we knew information about the variable. In real life it is often useful to work backwards from probabilities to estimate information about the data. This requires the **inverse normal distribution**.

KEY POINT 24.7

For a given value of probability p the inverse normal distribution gives the value of x such that $P(X \leq x) = p$.

EXAM HINT

Remember that p must be the *cumulative* probability.

EXAM HINT

Some calculators can find the value of x such that $P(X > x) = p$, as well as $P(X \leq x) = p$.

You will need to use your GDC to work out the inverse normal distribution (see Calculator skills sheet 14, the section on 'Finding the boundary' on the CD-ROM). To work out $P(X > x)$ you might need to do $1 - P(X \leq x)$. Note that many textbooks use the $\Phi(z)$ notation mentioned in the previous section to write inverse normal distribution: If $P(X \leq x) = p$, then $\Phi^{-1}(p) = z = \frac{x - \mu}{\sigma}$.

Worked example 24.7

The size of men's feet is thought to be normally distributed with mean 22 cm and variance 25 cm^2. A shoe manufacturer wants only 5% of men to be unable to find shoes large enough for them. How big should their largest shoe be?

Convert question into mathematical terms

If X is the crv 'length of a man's foot' then $X \sim N(22, 25)$

We want to find the value of x such that

$$P(X > x) = 0.05$$

Use inverse normal distribution. We may have to convert into a probability of the form $P(X \leq x)$

$P(X \leq x) = 1 - P(X > x) = 0.95$

$\Rightarrow x = 30.2$ cm (from GDC)

So their largest shoe must fit a foot 30.2 cm long.

EXAM HINT

This will involve solving equations, and sometimes simultaneous equations. As the numbers are usually not 'nice' you may want to use your calculator.

One of the main applications of statistics is to determine parameters of the population given information about the data. But how can we use the normal distribution calculations if the mean or the standard deviation is unknown? This is where the standard normal distribution comes in useful; we can replace all the X values by their Z-scores, as they follow a known distribution, N (0, 1).

Worked example 24.8

The masses of gerbils are thought to be normally distributed. If 30% of gerbils have a mass of more than 65 g and 20% have a mass of less than 40 g, estimate the mean and the variance of the mass of a gerbil.

continued . . .

Convert the information into mathematical terms

If X is the crv 'mass of a gerbil' then $X \sim N(\mu, \sigma^2)$

$$P(X > 65) = 0.3$$

$$P(X < 40) = 0.2 \quad (1)$$

If you need all the probabilities to be in the form $P(X \leq k)$, convert the first one

$$P(X \leq 65) = 0.7 \quad (2)$$

Use inverse normal distribution for Z ($Z \sim N(0, 1)$) and relate it to the given X values

from (1) $P(Z < z) = 0.2 \Rightarrow z = \dfrac{40 - \mu}{\sigma} = -0.842$

from (2) $P(Z \leq z) = 0.7 \Rightarrow z = \dfrac{65 - \mu}{\sigma} = 0.524$ (from GDC)

Solve simultaneous equations

$$40 - \mu = -0.842\sigma \quad (3)$$

$$65 - \mu = 0.524\sigma \quad (4)$$

(4) – (3)

$$25 = 1.366\sigma$$

$$\Rightarrow \sigma = 18.3g$$

$$\therefore \mu = 55.4g$$

Exercise 24D

1. (a) If $X \sim N(14, 49)$, find x if:

(i) $P(X < x) = 0.8$ (ii) $P(X < x) = 0.46$

(b) If $X \sim N(36.5, 10)$, find x if:

(i) $P(X > x) = 0.9$ (ii) $P(X > x) = 0.4$

(c) If $X \sim N(0, 12)$, find x if:

(i) $P(|X| < 0.5)$ (ii) $P(|X| < 0.8)$

2. (a) If $X \sim N(\mu, 4)$, find μ if

(i) $P(X > 4) = 0.8$ (ii) $P(X > 9) = 0.2$

(b) If $X \sim N(8, \sigma^2)$ find σ if

(i) $P(X \leq 19) = 0.6$ (ii) $P(X \leq 0) = 0.3$

3. If $X \sim N(\mu, \sigma^2)$, find μ and σ if:

(a) (i) $P(X > 7) = 0.8$ and $P(X < 6) = 0.1$

(ii) $P(X > 150) = 0.3$ and $P(X < 120) = 0.4$

(b) (i) $P(X > 0.1) = 0.4$ and $P(X \geq 0.6) = 0.25$

(ii) $P(X > 700) = 0.8$ and $P(X \geq 400) = 0.99$

4. IQ tests are designed to have a mean of 100 and a standard deviation of 20. What IQ score is needed to be in the top 2% of IQ scores? *[5 marks]*

5. Rabbits' masses are normally distributed with an average mass of 2.6 kg and a variance of 1.44 kg^2. A vet decides that the top 20% of rabbits are obese. What is the minimum mass for an obese rabbit? *[5 marks]*

6. A manufacturer knows that his machines produce bolts whose diameters follow a normal distribution with standard deviation 0.02 cm. He takes a random sample of bolts and finds that 6% of them have diameter greater than 2 cm. Find the mean diameter of the bolts. *[6 marks]*

7. (a) 30% of sand from Playa Gauss falls through a sieve with gaps of 1 mm, but 90% passes through a sieve with gaps of 2 mm. Assuming that a grain of sand's diameter is normally distributed, estimate the mean and standard deviation of the sand grains.

(b) 80% of sand from Playa Fermat falls through a sieve with gaps of 2 mm. 40% of this filtered sand passes through a sieve with gaps of 1 mm. Assuming that a grain of sand's diameter is normally distributed, estimate the mean and standard deviation of the sand grains. *[7 marks]*

8. The actual voltage of a brand of 9 V battery is thought to be normally distributed with standard deviation 0.8 V and mean $(9.2 - t)$ V where t is the time in hours that the battery has been used. When a battery's voltage drops below 7 V it can no longer power a lamp. A batch of batteries is found and only 10% can power the lamp. Assuming that the model is correct and that they were all used for the same amount of time, estimate for how long the batteries have been used. *[7 marks]*

9. The times taken for students to complete a test are normally distributed with a mean of 32 minutes and standard deviation of 6 minutes.

(a) Find the probability that a randomly chosen student completes the test in less than 35 minutes.

(b) 90% of students complete the test in less than t minutes. Find the value of t.

(c) A random sample of 8 students had their time for the test recorded. Find the probability that exactly 2 of

these students completed the test in less than 30 minutes. *[7 marks]*

10. An old textbook says that the range of data can be estimated as 6 times the standard deviation. If the data is normally distributed what percentage of the data is within this range? *[6 marks]*

11. A scientist noticed that 36% of temperature measurements were at least 4°C lower than the mean. Assuming that the measurements follow a normal distribution, estimate the standard deviation. *[5 marks]*

12. For a normal distribution find the ratios:

(a) $\dfrac{\text{median}}{\text{mean}}$

(b) $\dfrac{\text{standard deviation}}{\text{inter-quartile range}}$ *[6 marks]*

13. Evaluate $\Phi^{-1}(x)+\Phi^{-1}(1-x)$. *[3 marks]*

14. A company makes a large number of steel links for chains. They know that the force required to break any individual link is modelled by a normal distribution with mean 20 kN. The company tests chains consisting of 4 links. If any link breaks, the chain will break. A force of 18 kN is applied to all of the chains and 30% break.

(a) Estimate the probability of a single link breaking.

(b) Hence estimate the standard deviation in the breaking strength of the links. *[6 marks]*

15. Most calculators have a random number generator which generates random numbers distributed uniformly from 0 to 1. How can you use these to form random numbers that could be drawn from a normal distribution? *[4 marks]*

Summary

- Because we group continuous data, the probability of a **continuous random variable** (crv) is discussed in terms of the probability of it being in a given range. To do this we integrate a **probability density function** such that the area under the curve $f(x)$ represents the probability. The probability of the crv falling between values a and b is:

$$\mathrm{P}(a<x<b)=\int_a^b f(x)\,\mathrm{d}x$$